RESEARCH PAPER

# Ensemble Making Few-Shot Learning Stronger

Qiang Lin[1], Yongbin Liu[1,2†], Wen Wen[1], Zhihua Tao[1], Chunping Ouyang[1], Yaping Wan[1]

[1]Computer School, University of South China 42,1001, China

[2]Hunan provincial base for scientific and technological innovation cooperation, Hunan, China

**Keywords:** Few-shot learning; Relation extraction; Ensemble learning; Attention Mechanism; Fine-tuning

Citation: Lin, Q., et al.: Ensemble Making Few-Shot Learning Stronger. Data Intelligence 4(3), 529-551 (2022). doi: 10.1162/dint_a_00144

Received: Oct. 9, 2021; Revised: Jan. 15, 2022; Accepted: Feb. 12, 2022

## ABSTRACT

Few-shot learning has been proposed and rapidly emerging as a viable means for completing various tasks. Many few-shot models have been widely used for relation learning tasks. However, each of these models has a shortage of capturing a certain aspect of semantic features, for example, CNN on long-range dependencies part, Transformer on local features. It is difficult for a single model to adapt to various relation learning, which results in a high variance problem. Ensemble strategy could be competitive in improving the accuracy of few-shot relation extraction and mitigating high variance risks. This paper explores an ensemble approach to reduce the variance and introduces fine-tuning and feature attention strategies to calibrate relation-level features. Results on several few-shot relation learning tasks show that our model significantly outperforms the previous state-of-the-art models.

## 1. INTRODUCTION

Few-shot learning method is able to learn the commonness and specificity between tasks, and it can quickly and effectively generalize to new tasks by giving a few samples. The few-shot learning has become an approach of choice in many natural language processing tasks such as entity recognition and relation classification. There have been many few-shot models proposed in relation extraction tasks, such as siamese neural network [1], matching network [2], relation network [3], and prototypical network [4]. Among these models, the prototypical network is a more efficient and naive approach.

[†] Corresponding author: Yongbin Liu (E-mail: yongbinliu03@gmail.com; ORCID: 0000-0002-3369-3101).

However, representation learning and metrics selecting for few-shot relation classification are challenging due to the rich forms of relation expressions in natural language, usually containing local and global, complicated correlations between entities. These are also the leading cause of high variance problems [5]. It is problematic that a single model learns the representation for each relation well. Therefore, in few-shot scenarios, we propose a new ensemble method to improve the performance and the stability of the few-shot model. In the task of image classification, the ensemble method has been proposed by Dvornik et al. [6]. The method utilizes multiple identical sub-models based on resnet-18 and increases the diversity of sub-models by randomly knocking some sub-models off and transposing the image. However, this ensemble strategy requires resnet-18 with high complexity to adapt to the variation of the samples, and sub-models with a single structure is not applicable to text which is more complex and diverse than image. Considering the characteristic of text and the metric-based few-shot learning, we explore a new multi-structure ensemble model based on a prototypical network (MSEPN) in this paper, which aims to learn robust relation representations and similarity metrics from few-shot relation learning.

In the ensemble strategy of our MSEPN model, we integrate several high accuracies and diverse neural networks to learn the feature representations from each statement's semantics, rather than a sole network. These networks are designed for text and then they have been combined separately with multiple similarity metrics to measure the distance between texts in different sub-spaces. Compared with existing ensemble approaches, most of which utilize the same sub-models, we integrate this combination into the prototypical network to obtain sub-models with greater structural variability. The construction approach not only covers the shortage of each sub-model while taking their respective advantages but also helps the final ensemble model adapt to text which is diverse and complex using only a few layers of neural networks in sub-models. So far, we are the first to propose the diversity of neural network structures and similarity measures in ensemble strategy to increase the diversity of sub-models and adapt to the variability of textual expressions. Furthermore, in order to enhance the collaboration of the sub-models, we introduce adding a cooperation module with a joint loss function in our ensemble model MSEPN. The joint loss function is to combine the prediction losses of sub-models using a set of the learnable correlation coefficient for joint training, and then guide the models to learn a better way of cooperating, which can improve the overall performance of model prediction. Table 1 shows that the four improved prototypical models, which use different neural networks to learn relation and prototype representations respectively, perform on ten relation types where the similarity metric is adopted Euclidean distance and Cosine distance, respectively. As the result demonstrate, one sub-model specializes in one or more relation which is different from others. For example, Trans_Ed is good at recognizing P750 and P9 relation, while GRU_Cosine is good at P1411 and P7. After ensemble, our model performs best on almost all relations, where a collection of diverse representations often serves better together and can accommodate diverse text data containing a wider range of relation than a single strong one. Meanwhile, our strategy is able to avoid the effects of manual selection models.

**Table 1.** Performance of ten test relations on 5-way 5-shot relation classification task using prototypical networks. CNN denotes convolutional neural networks [7] encoder, Incep is inception networks [8] encoder, GRU is gated recurrent Unit networks [9] encoder, Trans is transformers networks [10] encoder, MSEPN is our ensemble approach, Ed is Euclidean distance, Cosine is cosine distance.

| Relation ID | CNN Ed | Incep Ed | GRU Ed | Trans Ed | CNN Cosine | Incep Cosine | GRU Cosine | Trans Cosine | MSEPN (Ensemble) |
|---|---|---|---|---|---|---|---|---|---|
| P750 | 88.24 | 87.97 | 88.83 | **90.62** | 83.89 | 89.16 | 83.31 | 89.23 | **92.43** |
| P1411 | 94.03 | 93.38 | 97.30 | 92.16 | 84.89 | 94.26 | **97.82** | 92.20 | 96.46 |
| P921 | **75.98** | 74.68 | 71.79 | 74.68 | 68.11 | 69.00 | 67.38 | 74.22 | **78.27** |
| P400 | 95.48 | 94.90 | 95.26 | 94.75 | 94.11 | **96.33** | 92.70 | 95.99 | **96.88** |
| P410 | 98.67 | 99.11 | **99.27** | 98.54 | 98.09 | 98.98 | 98.77 | 97.51 | **99.41** |
| P7 | 86.22 | 86.49 | 88.37 | 88.41 | 83.06 | 88.27 | **88.98** | 86.96 | **91.31** |
| P150 | 91.99 | 92.39 | 90.73 | 92.93 | 91.33 | **94.68** | 92.95 | 91.33 | 94.48 |
| P8 | 84.33 | **85.58** | 81.04 | 82.95 | 76.23 | 82.47 | 80.03 | 82.93 | **86.51** |
| P9 | 89.96 | 91.91 | 90.36 | **92.64** | 88.84 | 92.48 | 91.56 | 88.63 | **94.10** |
| P10 | 93.32 | 95.09 | 93.91 | 95.34 | 93.81 | **96.14** | 94.19 | 94.23 | 96.50 |

In order to further improve the domain adaptation and alleviate the prototype feature sparsity, we explore the fine-tuning and feature attention strategies to calibrate prototypical representations. The fine-tuning method was first proposed in the image classification field [5, 11, 12]. To adapt to the new relations, we use the cross-entropy loss function based on support samples to adjust the weights trained from scratch on an annotated corpus in our fine-tuning strategy. This strategy can significantly improve the accuracy of domain adaptation, especially in cross-domain few-shot relation extraction. In order to better learn the prototypical representation of each relation, we further propose feature attention to alleviate the problem of prototype feature sparsity. Motivated by the fact that, in each relation, the homogeneous features that are able to represent the relation between samples are closer in the feature space, we extract the proximity of homogeneous features from support samples of each relation, and the closer the feature, the higher the weights will be given. The attention mechanism can enhance classification performance and convergence speed.

We conduct experiments on the FewRel 1.0 dataset [13] extracting from wiki (without Domain Adaptation). Then, to show the ability of generalization, we choose PubMed, the domain of which is different from training (with Domain Adaptation). PubMed's data is collected from biomedical literature and manually annotated by FewRel 2.0 [14]. Experimental results demonstrate that our ensemble prototypical network significantly outperforms other baseline methods.

Our contributions consist of the following parts:

a) We propose a novel ensemble model MSEPN for few-shot relation extraction in the NLP domain. This method utilizes multiple neural networks and similarity metrics to fit in a wider range of relations, and a joint loss function to strengthen the collaboration of component models. Compared to a single model with a complex structure or other existing ensemble methods, MSEPN only requires a few simple models and is more suitable for few-shot tasks where the commonalities across different meta-tasks need to be learned.

b) To improve the similarity metric, we propose a feature attention method. The method is to exclude the useless features of the prototype and guide the predicted sample to be classified to the correct relation easier. It not only improves the performance of the few-shot model but also accelerates the convergence speed.

c) A new fine-tune method is designed for cross-domain few-shot relation extraction tasks which is more consistent with a realistic scenario. We introduce the MSEPN_FT model by combining the fine-tune method with MSEPN. MSEPN_FT makes full use of the limited samples in the support set, and it can alleviate the domain discrepancy problem which is a significant hindrance to generalizing to new tasks. The model achieves excellent performance on the tasks with more relations to be predicted. (The larger gap of the domain, the greater effect of fine-tuning.)

d) We propose the MSEPN_BERT model, which uses the pre-trained BERT as the word embedding for MSEPN. The model brings up the performance on the few-shot tasks with only a small number of relations that are to be classified.

## 2. RELATED WORK

In this section, we discuss the related work on few-shot learning.

**Parameters Optimization Learning:** In 2017, a meta-network based on gradient optimization has been proposed, which aims to utilize the learned knowledge and then rapidly generalize it to new tasks [15, 16]. A Meta network entirely relies on a base learner and a meta learner. The base learner gains the meta-information, which includes the information of the input task during dynamic representation learning, and then adaptively updates the parameters for the meta learner, while the meta learner memorizes the parameter and acquires the knowledge across different tasks [17, 18]. The agnostic method has been proposed by Finn et al. [19]. MAML looks for task distribution-sensitive initialization parameters in the hypothesis space that allows the model to be rapidly generalized to new tasks with only minor updates of a few samples. In further development, the few-shot optimization approach [16] not only looks for an initial parameter that is beneficial for fine-tuning but also adopts an LSTM-based optimizer to help the model adjust parameters. And then Bayesian Model Agnostic Meta-Learning [20, 21] proposed a new principled probabilistic framework, which utilizes scalable gradient-based meta-learning and nonparametric variational inference.

**Metric Based Few-Shot Learning:** Siamese neural network was applied to the few-shot classification by Koch et al. [1], and it utilized a convolutional architecture to rank the similarity between inputs naturally. Then, a matching network [2] was proposed in 2017. It used some external memories to enhance the neural networks. It added an attention mechanism and a new metric called cosine distance to predict the relations. MLMAN model [22] goes further improving the matching network. In 2018, a relation network is designed for few-shot learning by Sung et al. [3]. The relational network consists of two non-linear neural networks, one of which is used to learn the feature representation of the classes, and the other aims to measure the distance between the query sample and the support sample. Moreover, a large number of experiments show that Euclidean distance generally performs better than cosine distance in measuring the similarity between samples on multi-tasks. Thus, a simpler and more efficient model prototypical network was proposed by Snell et al. [4]. The naive approach used a standard Euclidean distance as the distance function. In 2019, Gao et al. [23] proposed a more efficient prototypical network, a hybrid attention-based prototypical network, and trained a weight matrix for Euclidean distance. These models depended on CNNs, RNNs, and Transformers [10] as the feature extractors. There are always some limitations for a single network to acquire semantic features.

**Fine-tuning Methods:** The idea of fine-tuning is used in few-shot learning, which refers to the pre-training model [24, 25, 26]. The fine-tuning deep network is a strong baseline for few-shot learning [11, 12]. These works connect the softmax cross-entropy loss with cosine distance. In 2020, Dhillon et al. [5] introduced a transductive fine-tuning baseline for few-shot learning. Most of these methods have been applied image domain, but are not widely explored in the relation extraction domain. Different from images, the text is more diverse and complicated. Therefore, we extend and generalize them to few-shot relation extraction tasks.

**Ensemble Method:** Ensemble Learning is effective in improving model performance and robustness. However, most of them have been applied in the CV or in large datasets. Yang et al. [27] proposed an ensemble model Ada-LSTMs for relation extraction in large sample scenarios. The method sequentially learns some weak classifiers and then votes to get the final prediction. Dvornik et al. [6] introduced the ensemble method to the task of image classification in few-shot scenarios. This strategy integrates multiple sub-model based on resnet-18 which has high time and space complexity and makes transformed images (cut, rotate, etc.) as the input of sub-models. During training, the method randomly knocks some sub-models off to increase the diversity of models. However, the single structure network limits the development and the extensibility of the model. Different from the ensemble strategy in the above two scenarios, we use different networks and metrics to create the sub-models, considering both the scenario and the characteristics of the metric-based model itself. Our strategy can overcome the shortcomings of the same structure and avoids manual selection of models to adapt to different relation texts.

In the paper, we discuss several factors that affect the robustness of few-shot relation learning. We propose the novel ensemble few-shot learning model MSEPN that integrates four networks and two metrics into a prototypical network. Furthermore, our proposed model adopts fine-tuning to improve the domain adaptation and feature attention strategy to address the problem of feature sparsity.

## 3. OUR APPROACH

In this section, we give a detailed introduction to the implementation of our proposed model MSEPN which is shown in Figure 1.

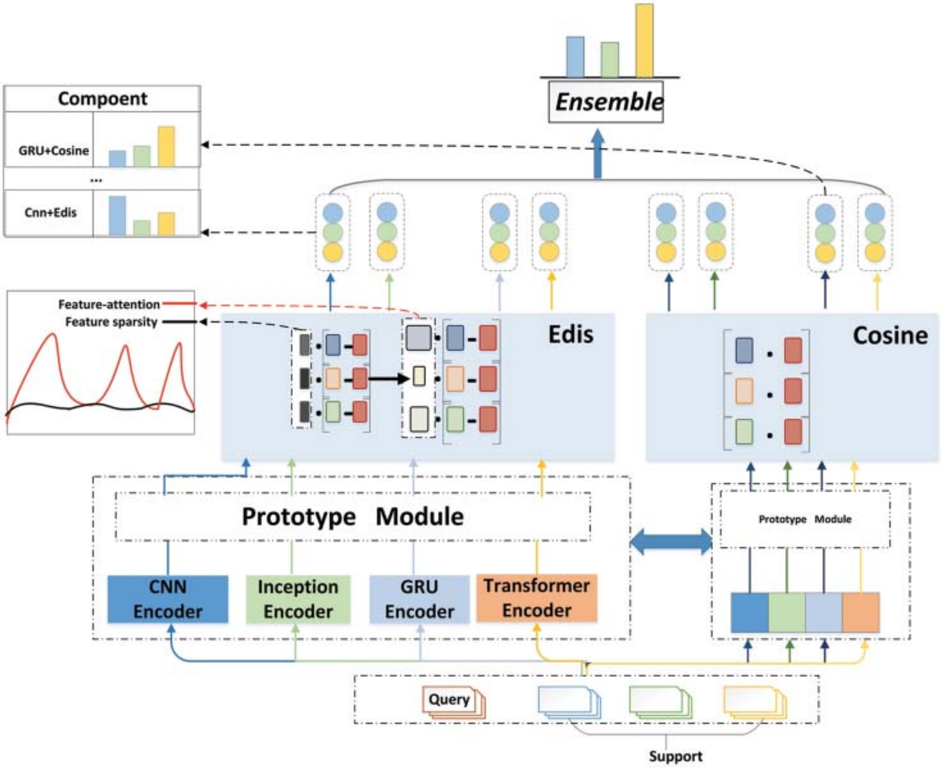


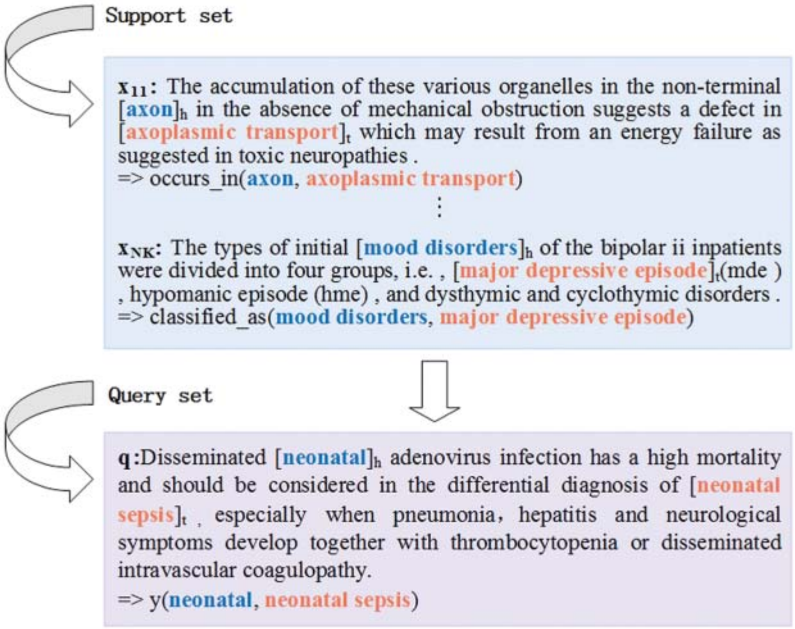
**Figure 1.** The architecture of our MSEPN model.

### 3.1 *Notations and Definitions*

We follow Gao et al. [23] and Snell et al. [4] to define our few-shot setting. Few-shot Relation classification (RC) is defined as a task to predict the relation $r$ between the entity pair $(h, t)$ mentioned in a query instance $x$. an example of a task is shown in Figure 2.

Given a relation $r \in R$ and a small support set of $N$ labeled examples $S = \{(x_1, y_1), \ldots, (x_{NK}, y_{NK})\}$, each $x_i \in \mathbb{R}^D$ is the $D$-dimensional feature vector of an example, and $y_i \in \{1, \ldots, NK\}$ is the corresponding label. As shown in Figure 2, $x$ is the labeled sample and $q$ is the query sample to be predicted. Few-shot relation extraction model predicts the relation $y$ of sample $q$ by seeing these labeled samples $S$.

The prototypical network [4] considers that there is a prototype representing a relation in the hypothetical space, in which the samples of each relation gather around their prototype, and the query samples are classified by calculating the closest prototype, which is obtained from samples in the support set. Given a sample $x = \{e_1, \ldots, e_n\}$ mentioning two entities, we encode the instance into a low-dimensional embedding $\mathbf{x} = f_\theta(x_i)$ through an embedding function with learnable parameters $\theta$, which are different for multiple neural



$x_{11}$: The accumulation of these various organelles in the non-terminal [axon]$_h$ in the absence of mechanical obstruction suggests a defect in [axoplasmic transport]$_t$ which may result from an energy failure as suggested in toxic neuropathies .
=> occurs_in(axon, axoplasmic transport)

⋮

$x_{NK}$: The types of initial [mood disorders]$_h$ of the bipolar ii inpatients were divided into four groups, i.e. , [major depressive episode]$_t$(mde ) , hypomanic episode (hme) , and dysthymic and cyclothymic disorders .
=> classified_as(mood disorders, major depressive episode)

q :Disseminated [neonatal]$_h$ adenovirus infection has a high mortality and should be considered in the differential diagnosis of [neonatal sepsis]$_t$ , especially when pneumonia, hepatitis and neurological symptoms develop together with thrombocytopenia or disseminated intravascular coagulopathy.
=> y(neonatal, neonatal sepsis)

**Figure 2.** An example of an N-way K-shot relation extraction task. $x_{11}$, …, $x_{NK}$ are the labeled samples. Occurs_in indicates the relation between head entity "axon" and tail entity "axoplasmic transport" in $x_{11}$, and classified_as is the relation label of $x_{NK}$. $y$ is the relation label of the sample $q$ that needs to be predicted.

networks encode layer. In our ensemble model, we adopt four classical neural networks to learn the **x** respectively. The main idea of prototypical networks is to compute a class representation $\mathbf{c}_k$ named prototype.

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} \mathbf{x}_i \tag{1}$$

Given a test point $x$, We can compute a distribution over classes as follow,

$$p_\theta(y = k \mid x) = \frac{exp(-d(\mathbf{x}, \mathbf{c}_k))}{\sum_{k'} exp(-d(\mathbf{x}, \mathbf{c}_{k'}))} \tag{2}$$

where $d$ is the distance function, which can be either Euclidean distance or Cosine distance. In our paper, both metrics are considered in our model. Also, we adopt feature-level attention to compute the $d$, which can achieve better results and convergence speed.

### 3.2 Feature Attention

The original model uses simple Euclidean distance as the distance function. In fact, some dimensions are more discriminative for classifying relations in the feature space [23]. We improve the feature-level attention in Gao et al. [23], and propose the feature attention based on vector subtraction that can find those dimensions more discriminated.

$$a_j = \sum_{i=1}^{n} \sum_{l=1, l \neq i}^{n} | x^i - x^l | / n^2 \tag{3}$$

where $n$ is the number of samples in the support set.

$$w_j = -\log\left(a_j + \xi\right) \tag{4}$$

where $\xi$ is a hyperparameter.

$$d_{ed}\left(x_q, \boldsymbol{c}_k\right) = w_k\left(x_q - \boldsymbol{c}_k\right)^2 \tag{5}$$

$$d_{cos}\left(x_q, \boldsymbol{c}_k\right) = w_k \frac{x_q^T \boldsymbol{c}_{\boldsymbol{k}}}{|| x_q ||_2 \cdot || \boldsymbol{c}_k ||_2} \tag{6}$$

where $w_k$ is the score vector for relation $r_{\boldsymbol{c}_k}$ computed via (3) and (4). The $d_{ed}$ refers to Euclidean distance and the $d_{cos}$ refers to Cosine distance.

### *3.3 Learning Ensemble of Deep Networks*

To reduce the high variance of few-shot learning, in MSEPN, we use ensemble methods acquiring semantic features, as in Figure 1. Now we discuss the objective functions of the learning ensemble prototypes model. During meta-training, each network needs to minimize the cross-entropy loss function over a training dataset:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, \gamma(f_\theta\left(x_i\right))\right) + \lambda\|\theta\|_2^2 \tag{7}$$

where $f_\theta$ is a basic neural network, which could be CNN, Inceptions, GRU, and Transformer in our experiments. $\gamma$ is a classifier with normalization. The cost function $\ell$ is the cross-entropy between label $y_i$ and $f_\theta(x_i)$, and the $\lambda$ is a weight decay parameter.

In our ensemble model, $E$ is the number of ensemble networks. During the training process, each network $f_{\theta_e}$ updates its parameters by minimizing its loss $L(\theta)$ on (7), $e \in E$. Due to choosing basic networks are very different, each network learns the semantic feature will typically differ. This is the reason of ensemble model appealing in our paper.

In our ensemble model, we propose a joint loss function to ensemble each network:

$$L(\tilde{\theta}) = \sum_{e=1}^{E}\left(\frac{1+w_e}{n}\sum_{i=1}^{n} \ell(y_i, \gamma(f_{\theta_e}(x_i))) + \lambda\left\|\theta_j\right\|_2^2\right), w_e \in \{w_1, \ldots, w_E\} \tag{8}$$

where $w_e$ is the correlation coefficient of $f_{\theta_e}$, we aim to seek outputs with a peaked posterior. We use a softmax layer to get the $w_e$. The loss function would help to solve the collaboration of the ensemble networks and leverage the variance of the classifiers.

### *3.4 Fine-tuning Learning*

For our model to recognize novel relation classes, especially for the cross-domain, we adopt a fine-tuning strategy to improve our model domain adaptation. We use a support set $S=\{\{x_i,y_i\}_{i=1}^{n_1},\ldots,\{x_i,y_i\}_{i=1}^{n_K}\}$ to fine-tuning our model parameters, where $n_k$ is the number of available samples of the class $k$. we encode $x_i$ of the $S$ to the representation and feed it to a feed-forward classifier (Softmax layer):

$$p_i = Softmax\left(\mathbf{M}\cdot f_{\tilde{\theta}}\left(x_i\right)+\mathbf{b}\right) \tag{9}$$

where **M** and **b** denote the learning parameters in the feed-forward layer, and we train it with the cross-entropy loss:

$$\min\sum_i CrossEntropy\left(y_i,p_i\right) \tag{10}$$

by (10), we could optimize the parameters of our model. By carefully initializing appropriately the parameter **M**, it is possible to achieve desirable properties of the ensemble [5]. So we use each class prototype in the $S$ to initialize **M**, setting **b** = **0**.

## 4. EXPERIMENTS

In this section, we present the experimental and implementation details of our proposed method. Then, we evaluate our proposed ensemble model MSEPN under two significant scenarios: cross-domain and in-domain, and compare our model with existing state-of-the-art (SOTA) models on different levels to demonstrate superiority and effectiveness. Finally, we further study the effect of each component that are utilized in our MSEPN model.

### *4.1 Datasets*

For various N-way K-shot setting, we evaluate our proposed model MSEPN on two open benchmarks: FewRel 1.0 [13] and FewRel 2.0 [28], which is presented in Table 2.

**Table 2.** Datasets.

| Source Dataset | Source | Apply | Relation # | Instance # |
|---|---|---|---|---|
| FewRel 1.0 [13] | Wiki | Training | 60 | 42,000 |
| | Wiki | Validation | 10 | 7,000 |
| | Wiki | Testing | 10 | 7,000 |
| FewRel 2.0 [28] | Wiki | Training | 64 | 44,800 |
| | SemEval-2010 task 8 | Validation | 17 | 8,851 |
| | PubMed | Testing | 10 | 2,500 |

FewRel 1.0 dataset has 100 relations and 700 examples for each. Due to the origin test set being hidden, we split the rest dataset into three parts: train, validation, and test set. To satisfy our experiment setting, we randomly choose 60 relations for the train set, 10 relations for the validation, and the rest 10 relations for testing. The relations of these three sets are disjoint, but they have the same distribution because all of them come from the Wikipedia corpus. However, it is not practicable enough for few-shot scenarios. Thus, we utilize FewRel 2.0 dataset, which is designed for cross-domain scenarios. The train set has 64 relations with 700 examples for each and comes from Wiki which is the same as FewRel 1.0. The validation set has 17 relations and is from SemEval-2010 task 8 which is annotated on news corpus, and then test on PubMed, which belongs to biomedical domains and has 10 relations. Moreover, we use the accuracy metric as the evaluation criteria.

### 4.2 Experimental Setup

The hyper-parameters of our model are shown in Table 3 and all the experiments are under the same hyper-parameters. We use the Glove 50-Dimensional word vectors as our initial word embeddings. In our model, the batch size is set to 4 with 30000 training iterations. During training, 20 relations are randomly selected in each batch, and then for each relation, both the number of support and query is set to 5. The optimizer of our model is stochastic gradient descent (SGD) with the weight decay of $10^{-5}$. We perform fine-tuning for 50 epochs and update the weight by the cross-entropy using support samples.

**Table 3.** Hyper-parameters Setting.

| | | | |
|---|---|---|---|
| Batch Size | 4 | Optimizer | SGD |
| N for train | 20 | Momentum | 0.9 |
| K for train | 5 | Weight Decay | 10-5 |
| Query Size | 5 | $\xi$ | 0.1 |
| Learning Rate | 0.1 | Training Iterations | 30000 |
| Val Step | 2000 | Fine-tune iterations | 50 |

### 4.3 Baselines

Siamese Network [1] maps the two input samples into a vector space, and then sends them into two identical sub-networks respectively. The method finds the similarity of the inputs by calculating the distance between its feature vectors.

FSL Graph Neural Networks (GNN) [29] is a typical method to measure a few-shot learning task with non-Euclidean data, the core idea of which is to disseminate information on a graph. In this model, the examples included in support and query are regarded as nodes in the graph, and then the information of the labeled support samples is passed to the query samples for relational inference by graph convolution.

Snail [30] is a meta-learning method that formalizes meta-learning as a sequence-to-sequence problem, using a combination of temporal convolution (TC) and attention mechanisms.

Prototypical Network (Proto) [4] believes that examples with the same relation label are close together and examples having different relations are far apart. Each query example can be classified by measuring the distance between itself and all prototypes of relations and then selecting the nearest prototype to acquire the relation label of this query.

Hybrid Attention-Base Prototypical Network (Proto_hatt) [23] is a variant of the prototypical network and addresses the noise problem. The method consists of an instance-level attention module that aims to select the most useful instances in the support set, and a feature-level attention module that is to select the most discriminate features for prediction. The method is the state-of-the-art few-shot models without pretraining.

BERT_PAIR [14] is a Bert-based sequential classification model. In this method, each query sample is concatenated with all support samples, and then the model is able to obtain a score for each pair of samples that represents the relation.

Resnet_Ensemble [6] is a few-shot image classification method based on ensemble learning. It utilizes 20 sub-models with resnet-18. Then it calculates the result by voting or averaging the output of multiple sub-models. During training, it increases the randomness of prediction by randomly eliminating several sub-models, and brings the variability up by giving each sub-model a transform of the same image.

### 4.4 Comparison Experiments with Existing methods

In this section, we show that our proposed models, including the ensemble model MSEPN, and its two enhanced versions (MSEPN_FT with our fine-tune strategy and MSEPN_Bert with pre-training) achieve better results on few-shot relation extraction under both cross-domain scenarios and in-domain scenarios. We further demonstrate that all the components we propose to integrate into our ensemble model, including ensemble method, feature attention mechanism, fine-tuning strategy, and pre-retraining approach, have a significant effect, and describe how they work in detail respectively.

We compare our proposed model with the typical models of few-shot relation extraction under a cross-domain (FewRel 2.0) scenario which makes sense for few-shot scenarios in reality. The comparison results are shown in Table 4. The table demonstrates the advantages of our proposed model from four aspects. First, when compared with the SOTA non-pre-trained model Proto_hatt, our model MSEPN has over 4% improvement for different scenarios. MSEPN_FT with our fine-tune strategy achieves higher performance and the maximum improvement is more than 22%. Moreover, our model MSEPN_BERT with pre-training has better performance than BERT-PAIR. Second, the comparison between our ensemble strategy and existing ensemble methods has obvious superiority. Resnet_Ensemble model is a typical few-shot ensemble model which is introduced to classify images by Dvornik et al. [6] and we transfer it to NLP. Since the sub-models of it share the same structure resnet-18, it requires high time and space, and its single structure is not suitable for handling text which is complex and diverse. To better accommodate the text, we improve its performance by replacing the resnet-18 which is the encoder with Transformer, and propose the Trans_Ensemble model. Though the performance of Trans_Ensemble is better than Resnet_Ensemble, our ensemble model MSEPN is the best because it is composed of diverse structures and designed for few-shot relation extraction. Third, the comparison of Proto_atten and Proto demonstrates the effect of our feature attention mechanism in our ensemble model. Proto_atten network utilizes the feature attention mechanism, while Proto is the naive prototypical network without the feature attention. As shown in Figure 3, Proto_atten improves a lot and the improvement is stable because the mechanism is able to discriminate the features which represent the relation.

**Table 4.** Results on cross-domain (FewRel2.0).

| Cross-domain (FewRel2.0) | 5 way | | | 10 way | | |
|---|---|---|---|---|---|---|
| | 1 shot | 5 shot | 10 shot | 1 shot | 5 shot | 10shot |
| Siamese | 39.66 | 47.72 | 53.08 | 27.47 | 33.58 | 38.84 |
| GNN | 35.95 | 46.57 | 52.20 | 22.73 | 29.74 | - |
| Proto | 40.16 | 52.62 | 58.69 | 28.39 | 39.38 | 44.98 |
| Proto_hatt | 40.78 | 56.81 | 63.72 | 29.26 | 43.18 | 50.36 |
| BERT-PAIR | 56.25* | 67.44* | - | 43.64* | 53.17* | - |
| Resnet_Ensemble | 38.50 | 51.93 | 57.95 | 26.77 | 38.90 | 44.77 |
| Trans_Ensemble | 41.59 | 57.33 | 62.85 | 29.80 | 43.63 | 49.34 |
| Proto_atten | 41.55 | 55.87 | 62.28 | 29.68 | 42.34 | 48.63 |
| MSEPN | 45.44 | 63.43 | 69.85 | 33.34 | 50.30 | 57.31 |
| MSEPN_FT | / | 72.40 | 78.00 | / | **65.31** | **71.60** |
| MSEPN_BERT | **58.30** | **76.22** | **80.12** | **44.05** | 62.71 | 67.22 |

Note: *Results reported by Gao et al. [28]. MSEPN_FT with fine-tune strategy. MSEPN_BERT with pre-training model Bert. / Our fine-tuning strategy is invalid for one-shot tasks, because it only has one support sample of each relation and cannot be formulated into few-shot scenario for our fine-tuning.

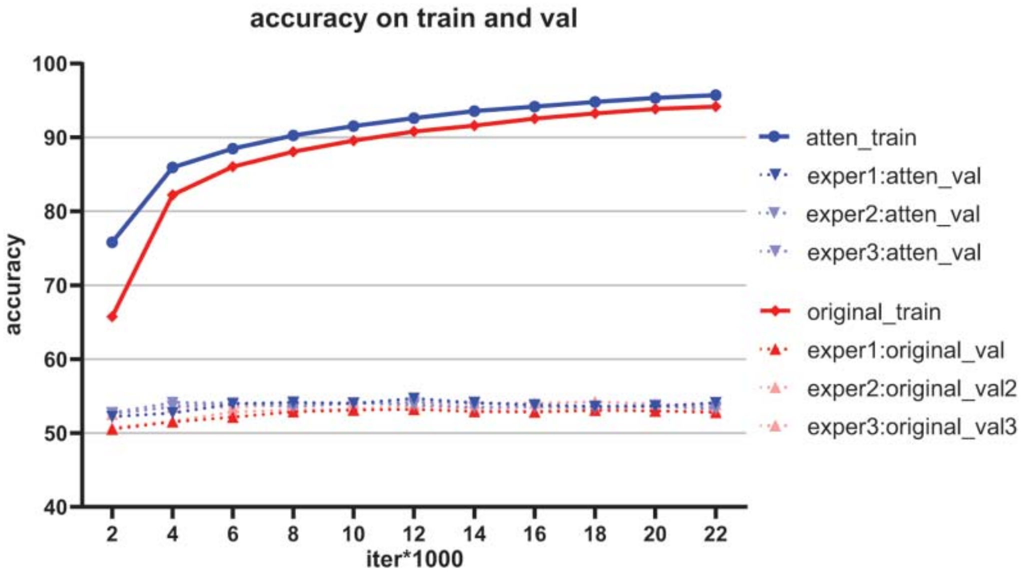


**Figure 3.** Comparison between proto_atten and proto.

Meanwhile, as this figure shows, Proto_atten achieves higher accuracy than Proto by nearly 20% when the training iteration is only 2000 and maintains the high value in future training, thus, the feature attention mechanism has an effect on accelerating convergence. Lastly, considering the novel fine-tuning strategy which is designed for cross-domain by us, our MSEPN_FT model with fine-tuning has a great improvement compared with MSEPN, and the performance of MSEPN_FT even higher than MSEPN_BERT with pretraining on complicated situations (10-way setting), since our fine-tuning strategy is able to make full of the small amount of support set to better adapt to the tasks in a new domain and improve the ability of generalization.

Since most of the existing methods are designed for the in-domain scenario, we also evaluate our proposed model on FewRel 1.0 dataset to compare fairness. The comparison results are shown in Table 5. On the whole, our proposed model MSEPN has a lot of improvement under all few-shot settings when compared with the existing SOTA non-pre-trained method. This table also demonstrates that our ensemble method, feature attention mechanism, fine-tuning strategy, and pre-training method play an important role in this task. MSEPN_BERT with pre-training achieves a higher level of performance, although MSEPN_FT with fine-tuning is less effective. Because the train set and test set share the same domain in the in-domain scenario, fine-tuning the support set of the test may aggravate the overfitting problem and lead to worse performance.

**Table 5.** Results on in-domain (FewRel 1.0). MSEPN_FT with fine-tune strategy. MSEPN_BERT with pre-training model Bert. / Our fine-tuning strategy is invalid for one-shot tasks because it only has one support sample of each relation and cannot be formulated into the few-shot scenario for our fine-tuning.

| In-domain (FewRel1.0) | 5 way | | | 10 way | | |
|---|---|---|---|---|---|---|
| | 1 shot | 5 shot | 10 shot | 1 shot | 5 shot | 10 shot |
| Siamese | 75.76 | 85.80 | 89.04 | 64.58 | 77.42 | 80.30 |
| GNN | 71.18 | 85.71 | 89.25 | 56.01 | 74.33 | - |
| Snail | 72.69 | 84.22 | 85.23 | 58.15 | 68.36 | 73.36 |
| Proto | 74.01 | 89.46 | 91.55 | 61.30 | 81.66 | 84.87 |
| Proto_hatt | 75.45 | 89.97 | 92.03 | 62.64 | 82.29 | 85.74 |
| Resnet_Ensemble | 74.29 | 88.56 | 90.88 | 61.08 | 80.20 | 83.72 |
| Trans_Ensemble | 78.01 | 89.63 | 91.56 | 66.71 | 82.30 | 85.21 |
| Proto_atten | 74.51 | 90.03 | 92.29 | 62.08 | 82.34 | 85.88 |
| MSEPN | 80.06 | 92.66 | 94.18 | 69.41 | 86.63 | 89.07 |
| MSEPN_FT | / | 92.73 | 94.12 | / | 87.46 | 89.59 |
| MSEPN_BERT | **86.08** | **95.25** | **96.14** | **78.28** | **91.07** | **92.52** |

### *4.5 Advantages of Ensemble Strategy*

In this part, we aim to show the advantages of our ensemble strategy from various aspects: model-level, task-level, and relation-level.

First of all, we study on the model level. We compare the performance of our proposed model Ensemble with its sub_models on a 5-way 5-shot relation extraction task, and the results under in-domain and cross-domain scenarios are shown in Figure 4. All of the sub-models in the Ensemble model exhibit significant fluctuations when testing on different datasets, for example, Incep_Ed model achieves the highest accuracy in the in-domain scenario while its performance is not satisfactory in the cross-domain scenario. However, our ensemble model which enables these sub-models to cooperate with each other and complement each strength is able to exceed the performance of the optimal sub-model and greatly improves the accuracy.

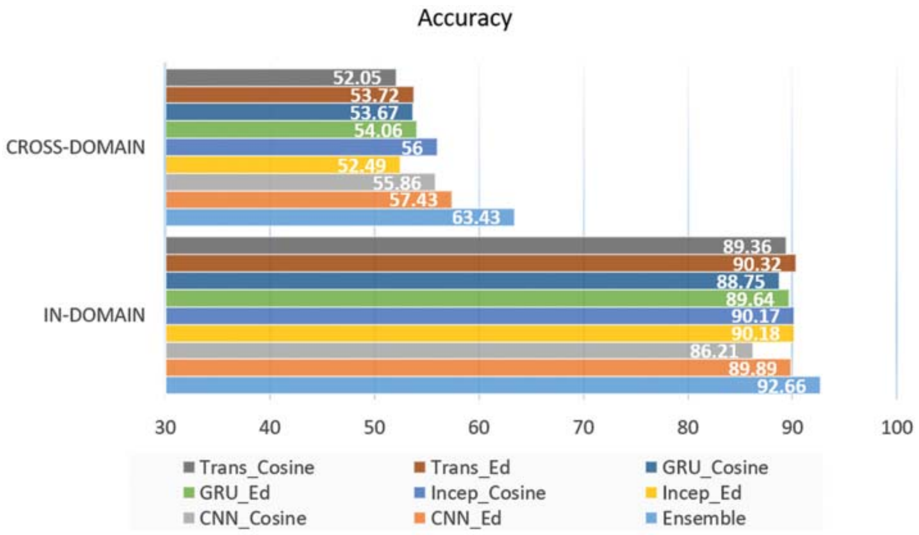


**Figure 4.** Comparison between MSEPN and sub-models on 5-way 5-shot relation extraction task.

Then, from the task-level analysis, the comparison between MS_Cosine and MS_Edis is shown in Table 6. MS_Cosine is a variant of our ensemble model in which sub-models only utilize the cosine as the similarity metric, while MS_Edis is the other variant that measures the similarity by Euclidean. As table 6 shows, both on 5-way 1-shot and 10-way 1-shot relation extraction tasks, MS_Cosine outperforms MS_Edis, while MS_Edis achieves higher accuracy on few-shot (K > 1) tasks. By integrating the advantages of Cosine and Edis, our ensemble model is able to improve the accuracy of each task and be more stable in the face of different settings.

**Table 6.** Comparison between MS_Cosine and MS_Edis. MS_Cosine is the ensemble model which is composed of CNN_Cosine, GRU_Cosine, Incep_Cosine, and Trans_Cosine, while the MS_Ed is composed of the rest sub-models with Ed.

| Domain | Model | 5 way | | | 10 way | | |
|---|---|---|---|---|---|---|---|
| | | 1 shot | 5 shot | 10 shot | 1 shot | 5 shot | 10 shot |
| Cross-Domain | MS_Cosine | **45.06** | 62.09 | 68.05 | **32.89** | 48.58 | 55.01 |
| | MS_Ed | 44.66 | **62.44** | **68.96** | 32.54 | 49.20 | 56.18 |
| | MSEPN | 45.44 | 63.43 | 69.85 | 33.34 | 50.30 | 57.31 |
| In-Domain | MS_Cosine | **79.80** | 91.93 | 93.55 | **69.00** | 85.43 | 87.95 |
| | MS_Ed | 79.44 | **92.52** | **94.03** | 68.60 | **86.49** | **88.96** |
| | MSEPN | 80.06 | 92.66 | 94.18 | 69.41 | 86.63 | 89.07 |

Last but not least, we demonstrate the advantages of our ensemble model from the relation level. In Figure 5, the sub-models have different performances on the examples with different relations. For example, CC which donates the CNN_Edis model is good at R2 and R4 relations, and GC which is the GRU_Cosine model is good at R1. However, our ensemble model achieves the best accuracy on all relations by integrating them.

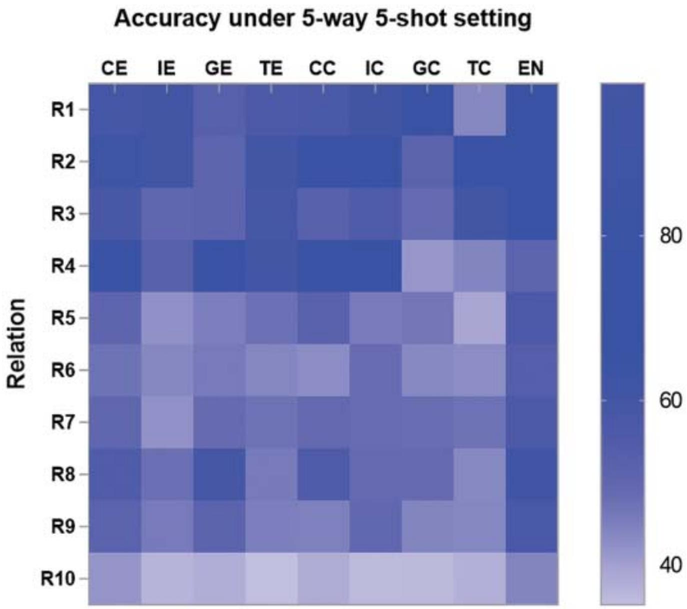


**Figure 5.** Accuracy of each relation in the testing set. The deeper color, the higher accuracy.

### *4.6 The stability of improvement*

In this part, we present the stability of our proposed model on FewRel 1.0 dataset and describe the reasons.

We randomly redistribute all relations in FewRel 1.0 dataset three times for supplement and then visualize the experiment results in Figure 6. All the evaluations have the same key point: When the number of examples in each relation is single, the performances of the Proto and Proto_hatt models drop dramatically, because the models are only based on prototypical networks which depend on the prototype of each relation is susceptible to noise data, while prototypical networks achieve higher accuracy with fewer parameters on few-shot tasks compared with other traditional approaches. Our ensemble strategy can solve this problem by the cooperation of the sub-models with different structures, and on few-shot tasks, our model inherits the advantage of the prototypical network and the parts including encoders and metrics help to recognize more diverse relations. Due to the characteristics demonstrate above, our ensemble model can not only improve the performance but also become more stable in all scenarios.

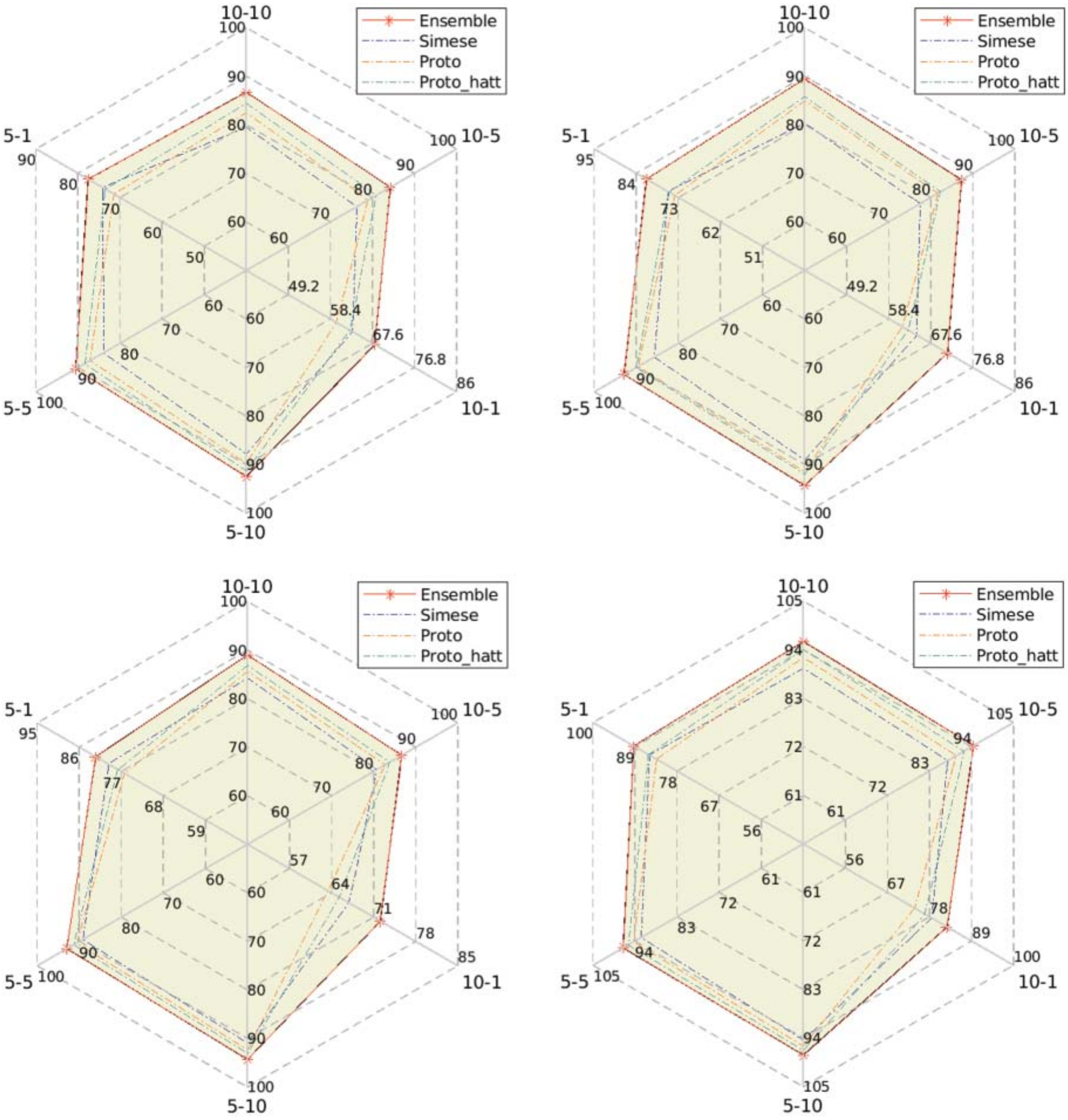


**Figure 6.** Radar plots on four Times random split Datasets between the performance of each model and the number of examples in each relation.

To further prove the stability of our ensemble model, we compare the four times experiments horizontally and then map the results to radar plots with four dimensions, which is shown in Figure 7. The closer the distance between each edge and the corresponding equipotential line (grey), the more stable the model is and the lower the variance. When compared to the other competitive models, our ensemble model achieves the highest accuracy in all of the four experiments, and the four connected edges of our model are very close to the equipotential lines, which demonstrates that our ensemble model can keep its superiority by ignoring the influence of different data partition. In terms of the specific value, the fluctuation ratio of our ensemble model has over 0.5% decrease compared with others, except the Snail network the accuracy of which is depressing on all tasks. Moreover, our ensemble model has more effective on datasets that are hard to predict. Because the relations are chosen randomly at each time, the above results are sufficient to prove that our method can leverage the variance no matter what relations are chosen in training, validation, or the testing set.

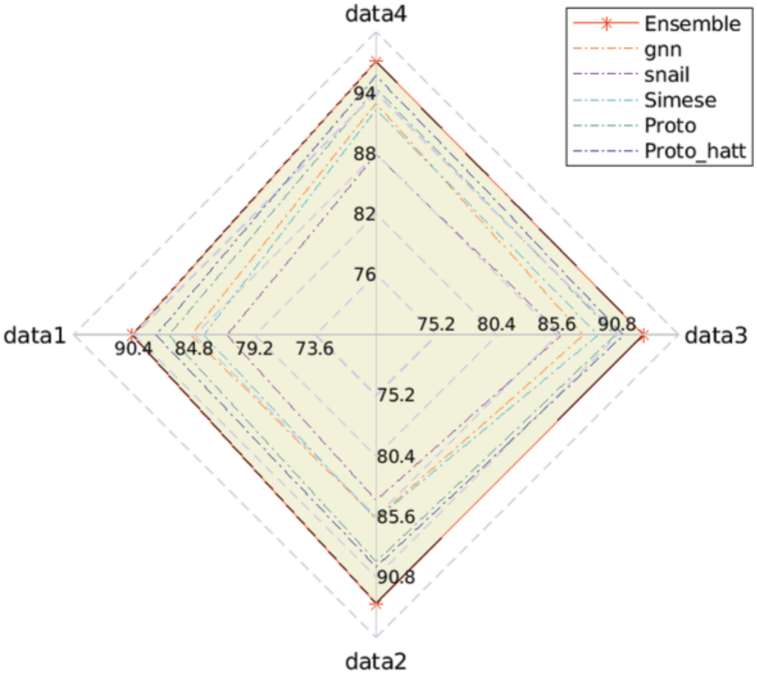


**Figure 7.** Radar plots on 4 times random split datasets to compare the stability of each model.

Above all, our ensemble model is able to improve the performance, reduce the variance, and enhance domain adaption. Our ensemble approach is necessary and effective for improving robustness.

### *4.7 Ablation Study*

In this section, we perform an ablation study to test the effectiveness of integrating metric and encoder, and the impact of the ensemble method and fine-tuning strategy.

*Effect of integrating various metrics*. We evaluate the two variants of our ensemble model. The first variant is MS_Ed in which the similarity metric utilizes only Euclidean distance, and the other variant is MS_Cosine with Cosine similarity. In a vector space, Euclidean similarity measures the real distance between points, while Cosine similarity focuses on the angle between the feature vectors. Since the two different measures of similarity have their function under different circumstances, the combination of them improves the performance of our ensemble model a lot, as Table 6 presents.

*Effect of integrating different structure encoders*. To show the effectiveness of integrating multiple encoders with different structures, we introduce four variants of our ensemble models named ‘En_CNN’, ‘En_Incep’, ‘En_GRU’, and ‘En_Trans’, in which the encoders are composed of four different neural networks (CNN, Inception, GRU, Transformer) respectively. The comparison is in Figure 8. Although En_Incep achieves the best performance under the in-domain scenario, the final ensemble model MSEPN can further improve the accuracy, while under the cross-domain scenario, En_CNN is the best. The above experiments demonstrate that our ensemble model which integrates different encoders is superior because the other encoders are able to supplement the model with high accuracy under both in-domain and cross-domain scenarios.

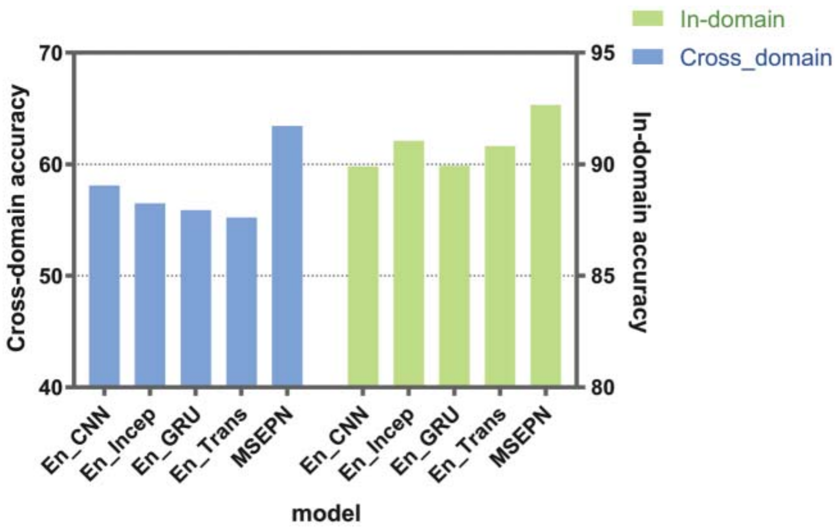


**Figure 8.** Accuracy of test data in 5-way 5-shot relation classification task in cross-domain and In-domain scenario. En_CNN denotes the ensemble model based on the CNN encoder that only integrates CNN_Ed with Euclidean similarity and CNN_Cosine with Cosine similarity. En_Incep, En_GRU, En_Trans are the ensemble models based on Inception, GPU, and Transformer respectively. MSEPN is our final ensemble model.

*Effect of the joint loss function*. In our ensemble method, we integrate a novel cooperative layer with joint loss function which adjusts the cooperative manner of sub-models by automatic learning the correlation coefficients. To demonstrate the effect of the cooperative layer, we compare it with the other method which direct concatenates the sub-models and does not have the cooperative layer. We experiment six times, and the results are presented in Figure 9. The accuracy points of our ensemble model MSEPN aggregate better, while the results of the other method fluctuate greatly and the value is twice ours because the cooperative layer in our model can guide and optimize the cooperative way of each sub-models. The above experiments fully illustrate our ensemble method can achieve higher performance and bring the robustness of our ensemble model up.

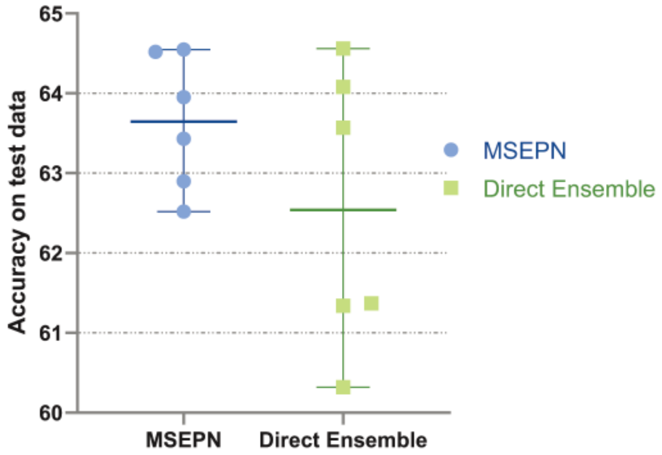


**Figure 9.** Comparison between different ensemble methods under cross-domain scenarios. The results of our method MSEPN are colored blue, while the model with the direct ensemble method is colored green. Each point is a result of an experiment. The shorter the line, the more stable the accuracy and the more robust the model.

*Effect of fine_tune strategy*. Our fine_tune strategy is designed to solve the problem of domain adaptation. In our proposed model MSEPN_FT, the initial parameter comes from MSEPN and the correlation coefficients of sub-models are learned automatically by the cooperative layer. The comparison between MSEPN and MSEPN_FT is shown in Figure 10. Our model MSEPN outperforms MSEPN_FT by over 9%, and MSEPN_FT is more robust. Moreover, our proposed model MSEPN_FT also achieves higher performance and is more stable than MSEPN_FT_RM which has been removed from the cooperative layer.

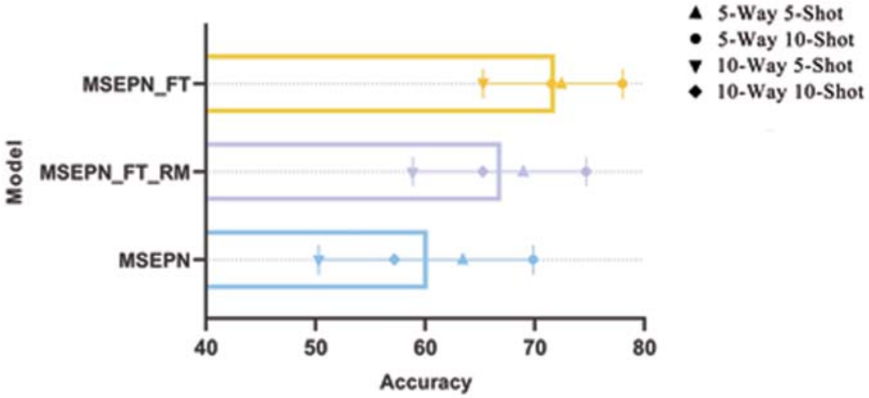


**Figure 10.** The effectiveness of fine-tuning. The rectangle is the average accuracy. Each point on the horizontal line is a result of an experiment. The shorter the horizontal line, the more stable the model.

## 5. CONCLUSION

In this paper, we propose an ensemble model MSEPN. Our MSEPN model consists of eight sub-models, which are based on prototypical networks with diverse neural networks and multiple similar metrics. We adopt fine-tuning for enhancing domain adaption and introduce feature attention which alleviates the problem of feature sparsity. In our experiments, we evaluate our model on FewRel 1.0 and FewRel 2.0, which demonstrate that our model significantly improves the accuracy and robustness, and achieves state-of-the-art results. In the future, we will explore more diverse ensemble schemes and adopt more neural encoders to make our model stronger.

## ACKNOWLEDGMENTS

The State Key Program of National Natural Science of China, Grant/Award Number: 61533018; National Natural Science Foundation of China, Grant/Award Number: 61402220; The Philosophy and Social Science Foundation of Hunan Province, Grant/Award Number: 16YBA323; Natural Science Foundation of Hunan Province, Grant/Award Number: 2020JJ4525, 2022JJ30495; Scientific Research Fund of Hunan Provincial Education Department, Grant/Award Number: 18B279, 19A439

## AUTHOR CONTRIBUTIONS

All authors contributed ideas, text, and review comments in the production of the paper. Q. Lin constructed and optimized the model. Q. Lin designed the experiment and analyzed the results.YB Liu proposed the core idea of the model and wrote the original draft. W.Wen participated in the model optimization and experimental design and analyzed the results. Z.H. Tao designed the structure of the model diagram and participated in the discussion of the results. C.P. Ouyang put forward the research topic and revised this paper.YP Wan provided important feedback and helped with the study.

## REFERENCES

[1] Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop, 2 (2015)

[2] Vinyals, O., Blundell, C., Lillicrap, T., et al.: Matching networks for one shot learning. Advances in Neural Information Processing Systems 29 (2016)

[3] Sung, F., Yang, Y., Zhang, L., et al.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1199–1208 (2018)

[4] Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning[J]. Advances in Neural Information Processing Systems 30 (2017)

[5] Dhillon, G.S., Chaudhari, P., Ravichandran, A., et al.: A baseline for few-shot image classification. arXiv preprint arXiv:1909.02729 (2019)

[6] Dvornik, N., Schmid, C., Mairal, J.: Diversity with cooperation: Ensemble methods for few-shot classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3723–3731 (2019)

[7] Le Cun, Y., Boser, B., Denker, J.S., et al.: Backpropagation applied to handwritten zip code recognition. Neural Computation 1(4), 541–551 (1989)

[8] Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)

[9] Cho, K., Van, Merriënboer, B., Gulcehre, C., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)

[10] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Advances in Neural Information Processing Systems 30 (2017)

[11] Chen, Y., Wang, X., Liu, Z., et al.: A new meta-baseline for few-shot learning. 2020

[12] Chen, W.Y., Liu, Y.C., Kira, Z., et al.: A closer look at few-shot classification. arXiv preprint arXiv:1904.04232 (2019)

[13] Han, X., Zhu, H., Yu, P., et al.: Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. arXiv preprint arXiv:1810.10147 (2018)

[14] Gao, T., Han, X., Zhu, H., et al.: FewRel 2.0: Towards more challenging few-shot relation classification. arXiv preprint arXiv:1910.07124 (2019)

[15] Munkhdalai, T., Yu, H.: Meta networks. In: International Conference on Machine Learning. PMLR, pp. 2554–2563 (2020)

[16] Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning (2016)

[17] Vanschoren, J.: Meta-learning: A survey. arXiv preprint arXiv:1810.03548 (2018)

[18] Elsken, T., Staffler, B., Metzen, J.H., et al.: Meta-learning of neural architectures for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12365–12375 (2020).

[19] Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. PMLR, pp. 1126–1135 (2017)

[20] Yoon, J., Kim, T., Dia, O., et al.: Bayesian model-agnostic meta-learning. Advances in Neural Information Processing Systems 31 (2018)

[21] Qu, M., Gao, T., Xhonneux, L.P., et al.: Few-shot relation extraction via bayesian meta-learning on relation graphs. In: International Conference on Machine Learning. PMLR, pp. 7867–7876 (2020)

[22] Ye, Z.X., Ling, Z.H.: Multi-level matching and aggregation network for few-shot relation classification. arXiv preprint arXiv:1906.06678 (2019)

[23] Gao, T., Han, X., Liu, Z., et al.: Hybrid attention-based prototypical networks for noisy few-shot relation classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 6407–6414 (2019)

[24] Devlin, J., Chang, M.W., Lee, K., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[25] Radford, A., Narasimhan, K., Salimans, T., et al.: Improving language understanding by generative pre-training (2018)

[26] Sarzynska-Wawer, J., Wawer, A., Pawlak, A., et al.: Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research 304, 114135 (2021)

[27] Yang, D., Wang, S., Li, Z.: Ensemble neural relation extraction with adaptive boosting. arXiv preprint arXiv:1801.09334 (2018)

[28] Gao, T., Han, X., Zhu, H., et al.: FewRel 2.0: Towards more challenging few-shot relation classification. arXiv preprint arXiv:1910.07124 (2019)

[29] Garcia, V., Bruna, J.: Few-shot learning with graph neural networks. arXiv preprint arXiv:1711.04043 (2017)

[30] Mishra, N., Rohaninejad, M., Chen, X., et al.: A simple neural attentive meta-learner[J]. arXiv preprint arXiv:1707.03141 (2017)

## AUTHOR BIOGRAPHY

**Qiang Lin** is studying for a master's degree in computer technology at the University of South China. Her research interests include Information Extraction and Few-shot learning.

**Yongbin Liu** received a Ph.D. degree from the University of Science & Technology Beijing, China, in 2013. From 2013 to 2015, he was a post-doc research fellow at Tsinghua University. He is an associate professor at the University of South China. His research interests include natural language processing and knowledge engineering.

**Wen Wen** received his M.E degree from the University of South China, China, in 2021. Her research interests focus on Relation Extraction and Few-shot learning

**Zhihua Tao** is studying for a master's degree in computer technology at University of South China. Her research interests include information extraction and knowledge mapping.

**Chunping Ouyang** received a Ph.D. degree from the University of Science & Technology Beijing, China, in 2011. She is a professor of computer science at the University of South China and supervisor of postgraduate. Her research interests include natural language processing and information retrieval.

**Yaping Wan** received a Ph.D. degree from the Huazhong University of Science and Technology, China, in 2009. He is a professor of computer science at the University of South China and supervisor of postgraduate. His research interests include data structure and information retrieval.